Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

Null- und Alternativhypothese

Die statistischen Hypothesen beziehen sich auf einen unbekannten Parameter ϑ der Verteilung (des Modells) des interessierenden Merkmals in der unterliegenden Population

Als Nullhypothese H_0 wird üblicherweise eine Hypothese der folgenden Art formuliert

$$\vartheta = \vartheta_0 \quad \text{oder} \quad \vartheta \leq \vartheta_0 \quad \text{oder} \quad \vartheta \geq \vartheta_0$$

Diesen Nullhypothesen stehen folgende Alternativhypothesen H_1 gegenüber

H_0	H_1	
$\vartheta = \vartheta_0$	$\vartheta \neq \vartheta_0$	(zweiseitig)
$\vartheta \leq \vartheta_0$	$\vartheta > \vartheta_0$	(einseitig)
$\vartheta \geq \vartheta_0$	$\vartheta < \vartheta_0$	(einseitig)

Prüfgröße

Die Überprüfung der statistischen Hypothesen erfolgt dann über die Ziehung einer (einfachen) Zufallsstichprobe $X_1, \ldots, X_n$ vom Umfang n, die eine Realisierung $x_1, \ldots, x_n$ liefert

Ausgehend von einem Stichprobenkennwert sollen dann Inferenzen bezüglich der Populationsparameter gezogen werden, es wird also ein Schätzwert $\hat{\vartheta}$ für den Parameter ϑ berechnet

$$\hat{\vartheta} = t(x_1, \ldots, x_n)$$

Die entsprechende Schätzfunktion

$$T = t(X_1, \ldots, X_n)$$

wird auch als *Teststatistik* bzw. *Prüfgröße* bezeichnet

Stichprobenverteilung

Unter der Voraussetzung, dass ϑ der wahre Parameter ist, interessiert man sich nun für die Wahrscheinlichkeit, mit der bestimmte Werte der Prüfgröße auftreten, d. h. für bedingte Wahrscheinlichkeiten, wie etwa

$$P(c_1 \leq T = t(X_1, \ldots, X_n) \leq c_2 \mid \vartheta)$$
$$P(T = t(X_1, \ldots, X_n) \leq c \mid \vartheta)$$

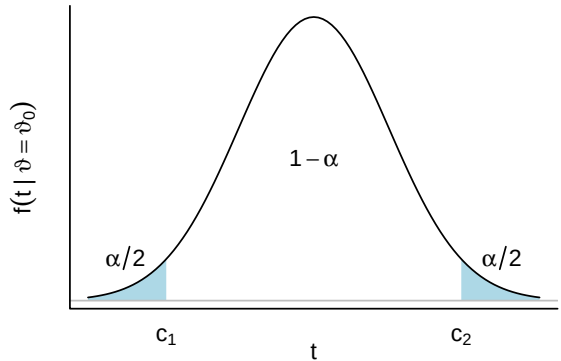
Hierzu wird bei parametrischen Tests unter der Annahme der Gültigkeit der Nullhypothese H_0 aus der Verteilung in der Population die Verteilung der Zufallsvariablen T abgeleitet

- Häufig wird als Verteilung in der Population eine Normalverteilung angenommen, was sich auf eine Reihe vorteilhafter Eigenschaften gründet (die wichtigsten haben wir bereits besprochen)

Nullhypothese $H_0\colon \vartheta = \vartheta_0$ und Alternativhypothese $H_1\colon \vartheta \neq \vartheta_0$

Bei vorgegebener Wahrscheinlichkeit α des Fehlers 1. Art kann man Konstanten c_1 und c_2 so bestimmen, dass gilt

$$P(T < c_1 \mid \vartheta = \vartheta_0) = P(T > c_2 \mid \vartheta = \vartheta_0) = \frac{\alpha}{2}$$



$H_0 \colon \vartheta = \vartheta_0$ und $H_1 \colon \vartheta \neq \vartheta_0$

Testentscheidung

- Gilt $t(x_1, \ldots, x_n) < c_1$ oder $t(x_1, \ldots, x_n) > c_2$, so wird H_0 abgelehnt (Ablehnungsbereich)
- Für $c_1 \leq t(x_1, \ldots, x_n) \leq c_2$ wird H_0 nicht abgelehnt

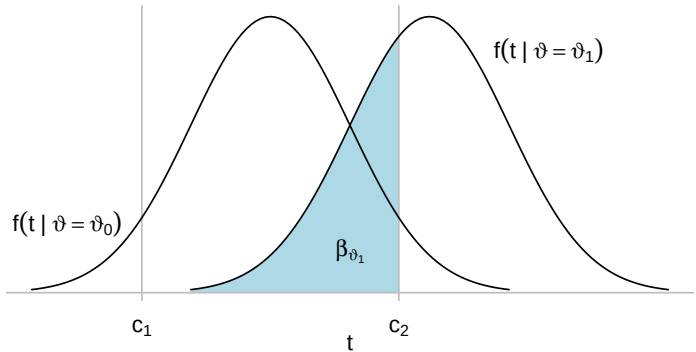
Ein Fehler 2. Art wird begangen, wenn das Ereignis $c_1 \leq T \leq c_2$ eintritt und gleichzeitig für den wahren Parameter $\vartheta = \vartheta_1 \neq \vartheta_0$ gilt

Für die Wahrscheinlichkeit des Fehlers 2. Art unter der Annahme $\vartheta = \vartheta_1$ gilt dann

$$\beta_{\vartheta_1} = P(c_1 \leq T \leq c_2 \mid \vartheta = \vartheta_1)$$

$H_0: \vartheta = \vartheta_0$ und $H_1: \vartheta \neq \vartheta_0$

Grafische Darstellung des Fehlers 2. Art unter der Annahme $\vartheta = \vartheta_1$



$H_0\colon \vartheta = \vartheta_0$ und $H_1\colon \vartheta \neq \vartheta_0$

Wenn die Nullhypothese H_0 nicht zutrifft, dann kann jede der Zahlen $\vartheta \neq \vartheta_0$ als Wert des Parameters in Betracht kommen

Daher betrachtet man für die Bestimmung der Wahrscheinlichkeit β des Fehlers 2. Art das Supremum (die kleinste obere Schranke) der entsprechenden Wahrscheinlichkeiten

$$\beta = \sup_{\vartheta \neq \vartheta_0} P(c_1 \leq T \leq c_2 \mid \vartheta)$$

(d. h. β ist größer oder gleich als alle Wahrscheinlichkeiten auf der rechten Seite und ist die kleinste Zahl mit dieser Eigenschaft)

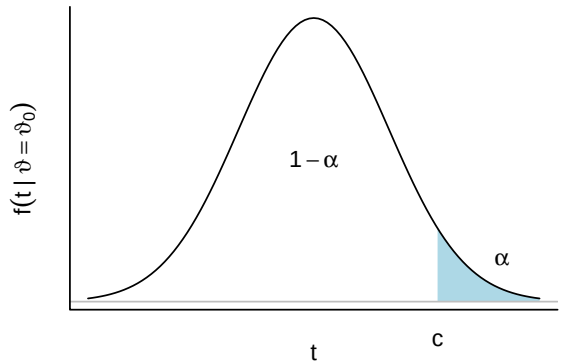
Wird also kein minimal relevanter Effekt festgelegt, dann

- nimmt β den Wert $1 - \alpha$ an und beträgt die Teststärke lediglich α
- sind somit keine gültigen Schlussfolgerungen möglich

Nullhypothese $H_0\colon \vartheta \leq \vartheta_0$ und Alternativhypothese $H_1\colon \vartheta > \vartheta_0$

Bei vorgegebener Wahrscheinlichkeit α des Fehlers 1. Art kann man bei den meisten Problemstellungen eine Konstante c so bestimmen, dass gilt

$$\alpha = \sup_{\vartheta \leq \vartheta_0} P(T > c \mid \vartheta)$$



$H_0 \colon \vartheta \leq \vartheta_0$ und $H_1 \colon \vartheta > \vartheta_0$

Testentscheidung

- Gilt $t(x_1, \ldots, x_n) > c$, so wird H_0 abgelehnt (Ablehnungsbereich)
- Für $t(x_1, \ldots, x_n) \leq c$ wird H_0 nicht abgelehnt

Die Bestimmung der Wahrscheinlichkeit des Fehlers 2. Art unter der Annahme $\vartheta = \vartheta_1 > \vartheta_0$ erfolgt durch

$$\beta_{\vartheta_1} = P(T \leq c \mid \vartheta = \vartheta_1)$$

Ohne Festlegung eines minimal relevanten Effekts würde gelten

$$\beta = \sup_{\vartheta > \vartheta_0} P(T \leq c \mid \vartheta)$$

Gütefunktion

Die bei einem statistischen Test auftretenden Wahrscheinlichkeiten der Fehler 1. und 2. Art kennzeichnen die Güte dieses Tests, die durch die *Gütefunktion* $g(\vartheta)$ beschrieben wird

Man betrachtet die bedingte Wahrscheinlichkeit, die Nullhypothese H_0 abzulehnen gegeben den Parameterwert ϑ

$$g(\vartheta) = P(H_0 \text{ ablehnen} \mid \vartheta)$$

Es ist

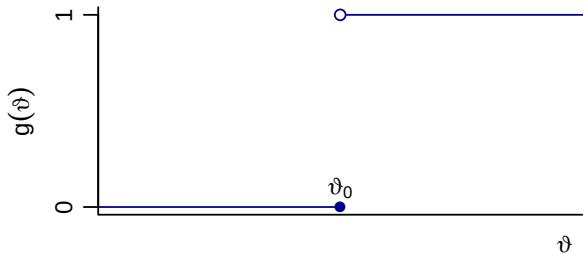
$$\begin{aligned} P(\text{Fehler 1. Art}) &= P(H_0 \text{ ablehnen} \mid H_0 \text{ wahr}) \\ P(\text{Fehler 2. Art}) &= P(H_0 \text{ beibehalten} \mid H_1 \text{ wahr}) \\ &= 1 - P(H_0 \text{ ablehnen} \mid H_1 \text{ wahr}) \end{aligned}$$

Ideale Gütefunktion

Idealerweise sollte die Gütefunktion $g(\vartheta)$ den Wert 0 annehmen, wenn ϑ aus dem Bereich der Nullhypothese H_0 stammt, und 1, wenn ϑ im Bereich der Alternativhypothese H_1 liegt

Für einen einseitigen Test mit Nullhypothese $H_0\colon \vartheta \leq \vartheta_0$ und Alternativhypothese $H_1\colon \vartheta > \vartheta_0$ gilt dann im Idealfall

$$g(\vartheta) = \begin{cases} 0 & \text{für } \vartheta \leq \vartheta_0 \quad (\vartheta \in H_0) \\ 1 & \text{für } \vartheta > \vartheta_0 \quad (\vartheta \in H_1) \end{cases}$$



Gütefunktion, α und β

Eine solche Gütefunktion $g(\vartheta)$ gehört zu einem Test, in dem weder ein Fehler 1. Art noch ein Fehler 2. Art auftritt, was praktisch nicht erreichbar ist

Da der statistische Test die Irrtumswahrscheinlichkeit α kontrolliert, gilt für $\vartheta \in H_0$

$$g(\vartheta) \leq \alpha$$

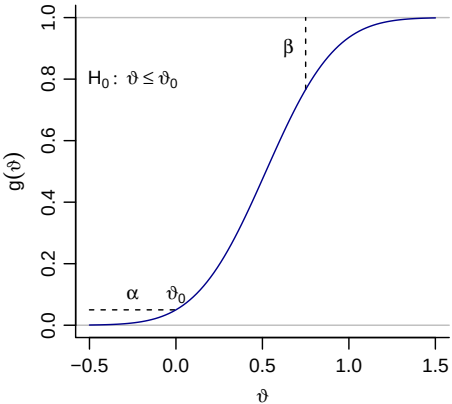
Für $\vartheta \in H_1$ gilt

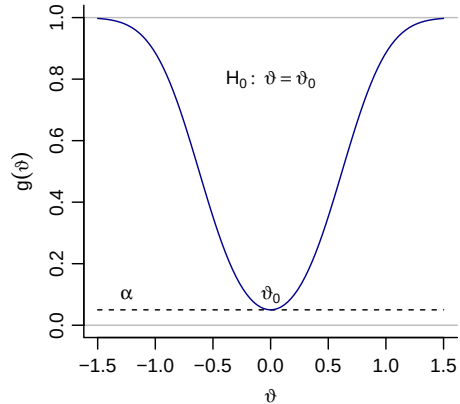
$$g(\vartheta) = 1 - \beta_\vartheta$$

(Teststärke oder Power) mit der Wahrscheinlichkeit β_ϑ eines Fehlers 2. Art für den Parameterwert ϑ unter Gültigkeit der Alternativhypothese

Gütefunktion, α und β

Die folgende Abbildung zeigt typische Gütefunktionen





Gütefunktion und Teststärke

Für Werte aus H_0 ist die Gütefunktion kleiner gleich α

Für Werte aus H_1 heißt die Gütefunktion *Teststärke* (auch *Power* oder *Trennschärfe*)

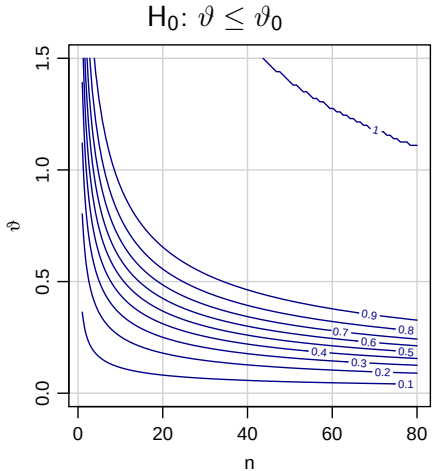
Die Teststärke wächst mit dem Stichprobenumfang n, d. h. die Gütefunktion wird steiler

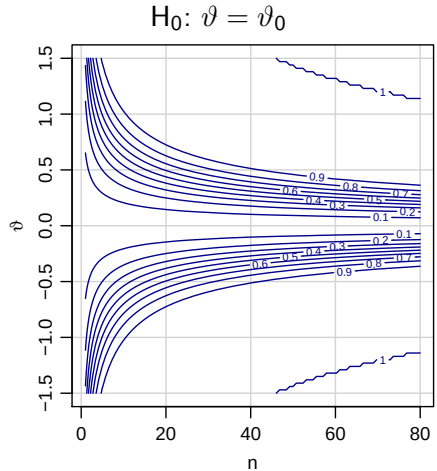
Die Teststärke wächst mit zunehmender Irrtumswahrscheinlichkeit α

Für eine wachsende Abweichung zwischen Werten aus H_1 und H_0 wird die Teststärke größer

Gütefunktion und Teststärke

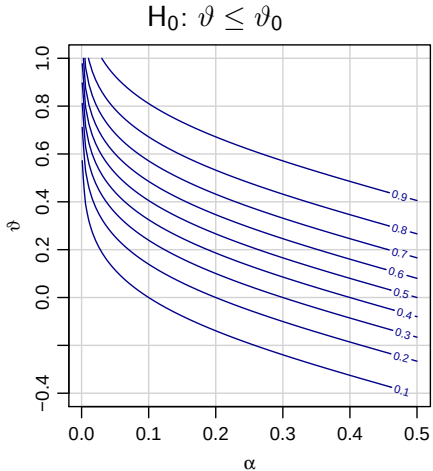
Die Teststärke wächst mit dem Stichprobenumfang n ($\vartheta_0 = 0$)

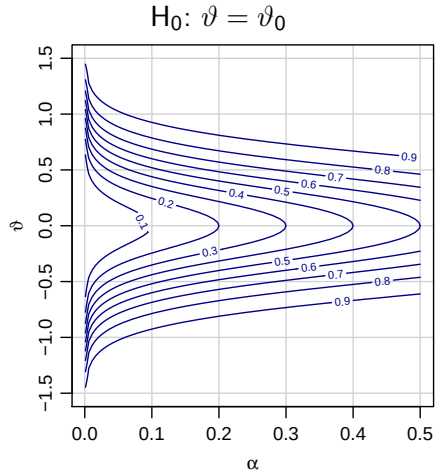




Gütefunktion und Teststärke

Die Teststärke wächst mit zunehmender Irrtumswahrscheinlichkeit α ($\vartheta_0 = 0$)





p-Werte

Das Ergebnis statistischer Tests wird häufig auch durch *p-Werte* oder *Überschreitungswahrscheinlichkeiten* beschrieben

- Die p-Werte werden durch Statistiksoftware (z. B. R, SPSS) standardmäßig ausgegeben

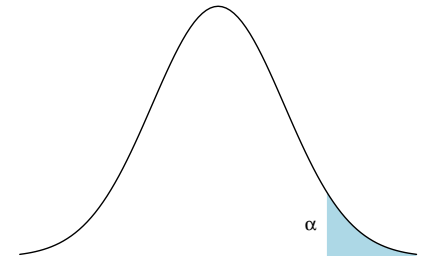
Ein p-Wert gibt die Wahrscheinlichkeit an, unter der Gültigkeit der Nullhypothese H_0 einen Wert der Prüfgröße zu erhalten, der mindestens so extrem (in Richtung der Alternativhypothese) ist, wie der für die Stichprobe berechnete Wert

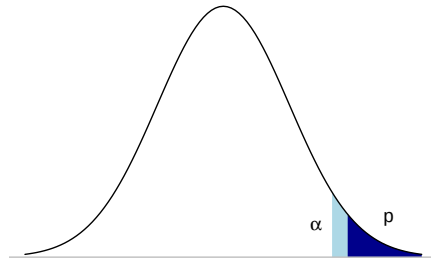
Ein sehr kleiner p-Wert bedeutet, dass es unter H_0 sehr unwahrscheinlich ist den Wert der Prüfgröße (bzw. einen noch extremeren) zu beobachten

- Dies wird als Evidenz gegen die H_0 gewertet
- H_0 wird verworfen, wenn der p-Wert kleiner als α ist

Einseitiger Test

Vergleich mit Irrtumswahrscheinlichkeit α bei einseitigem Test





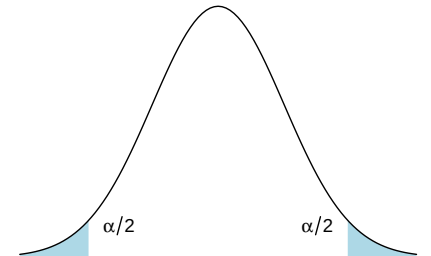
Zweiseitiger Test

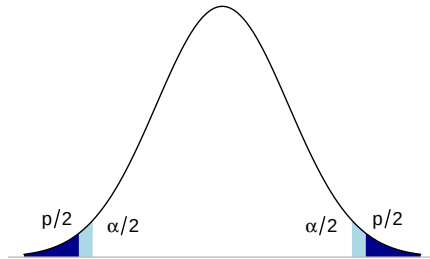
Vergleich mit Irrtumswahrscheinlichkeit α bei zweiseitigem Test

- Bei einer zweiseitigen Fragestellung verteilt sich der als Bereich der Gültigkeit der Alternativhypothese angesehene Ablehnungsbereich auf beide "Enden" der Verteilung unter der Nullhypothese H_0
- Daher können Abweichungen von dem in H_0 postulierten Wert in zwei Richtungen auftreten
- Neben den Werten der Prüfgröße, die in Richtung auf den für die Stichprobe berechneten Wert mindestens so extrem abweichen, sind auch die in der Gegenrichtung von der H_0 mindestens so extrem abweichenden Werte der Prüfgröße zu betrachten

Zweiseitiger Test

Vergleich mit Irrtumswahrscheinlichkeit α bei zweiseitigem Test





p-Werte bei Fisher und Neyman-Pearson

Die durch den Vergleich mit α suggerierte Interpretation der p-Werte als Irrtumswahrscheinlichkeiten ist problematisch

Die Probleme resultieren aus einer Vermengung zweier unterschiedlicher Traditionen des statistischen Testens

- Die ursprünglich von *R. A. Fisher* entwickelte Theorie des Signifikanztests
- Die heute als Standard geltende Theorie des statistischen Hypothesentests nach *Jerzy Neyman* und *Egon Pearson*

Der Begriff des p-Werts stammt aus der Theorie von Fisher, während der Begriff der Irrtumswahrscheinlichkeit α aus der Theorie von Neyman-Pearson stammt

Exkurs: Signifikanztest nach Fisher

Fisher (1925, 1935) geht von der Formulierung einer Nullhypothese aus (der nicht wie bei Neyman-Pearson eine Alternativhypothese gegenüber steht), die annimmt, dass eine Stichprobe aus einer Population mit bekannter Stichprobenverteilung stammt

Die Nullhypothese wird als abgelehnt ("disproved") betrachtet, wenn die Teststatistik vom Erwartungswert der Stichprobenverteilung um mehr als einen bestimmten Kriteriumswert abweicht, dem Signifikanzniveau (p-Wert)

> *"Fisher regarded p values as constituting inductive evidence against the null hypothesis; the smaller the p value, the greater the weight of said evidence (Johnstone 1986, 1987b; Spielman 1974). In terms of his famous disjunction, a p value $\leq$ .05 on the null hypothesis indicates that 'either an exceptionally rare chance has occurred or the theory is not true' (Fisher 1959, p. 39)." (Hubbard & Bayarri, 2003)*

Exkurs: Signifikanztest nach Fisher

Fisher betrachtete die Berechnung von p-Werten als "objektive" Methode, mit der ein Forscher die Plausibilität der Nullhypothese prüfen kann

> *"... the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to and verifiable by other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief [in the null hypothesis] it engenders." (Fisher, 1959, S. 43; zit. nach Hubbard & Bayarri, 2003)*

Die Festlegung des Signifikanzviveaus kommentiert Fisher folgendermaßen

> *"A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance. ... If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point)." (Fisher, 1926, S. 504; zit. nach Hubbard & Bayarri, 2003)*

Exkurs: Hypothesentest nach Neyman-Pearson

Die Theorie von Neyman-Pearson (1928, 1933) wurde als Weiterentwicklung des Ansatzes von Fisher nach dem 2. Weltkrieg populär

Durch die Verwendung des Begriffs "Hypothesentest" setzen sie sich gegen den Signifikanztest nach Fisher ab

Die Grundlage bildet die Formulierung zweier komplementärer Hypothesen, der Nullhypothese H_0 und der Alternativhypothese H_1

Zur Betrachtung der mit einer Entscheidung für diese Hypothesen verbundenen Fehler werden die Wahrscheinlichkeiten α der falschen Ablehnung (Fehler 1. Art) und β für die falsche Beibehaltung (Fehler 2. Art) eingeführt

- Diese Wahrscheinlichkeiten werden im Kontext einer wiederholten Ziehung von Zufallsstichproben frequentistisch interpretiert (vgl. Konfidenzintervalle)

Exkurs: Hypothesentest nach Neyman-Pearson

Der Ansatz von Neyman-Pearson führt den vollständig neuen Begriff der Teststärke (Power) ein, der durch die Wahrscheinlichkeit $1 - \beta$ gegeben ist, eine falsche Nullhypothese zu verwerfen

Die Stärke eines Tests, eine bestimmte *Effektgröße* in der Population zu entdecken, wird vor jeder Datenerhebung berechnet und ist eine entscheidende Größe für die experimentelle Versuchsplanung

Auch die Irrtumswahrscheinlichkeit α ist vor der Datenerhebung festzulegen, der konkrete Wert richtet sich nach inhaltlichen Maßstäben (per Konvention meist $\alpha = 0.05$)

Diese Charakteristika beziehen sich daher nicht auf die Durchführung eines einzelnen Tests, sondern auf eine ganze Klasse von Tests (basierend auf einer wiederholten Zufallsauswahl)

p-Wert und Irrtumswahrscheinlichkeit α

Der datenabhängige p-Wert im Ansatz von Fisher ist daher klar zu unterscheiden von der a priori festgelegten Irrtumswahrscheinlichkeit α

- Der p-Wert bezieht sich auf ein einzelnes Experiment

In der Neyman-Pearson-Tradition kann ein p-Wert lediglich als Prüfgröße betrachtet werden, mit der die Information in den Daten zusammengefasst wird

- Dieser Wert kann dann mit einer a priori festgelegten Irrtumswahrscheinlichkeit α verglichen werden

Die Angabe einer post hoc bestimmten Irrtumswahrscheinlichkeit zu einem resultierenden p-Wert (wird durch Ausgabe in Statistiksoftware nahegelegt) ist im Neyman-Pearson-Ansatz unzulässig

Beziehung zu Konfidenzintervallen

Die Entscheidung in einem Hypothesentest kann z. B. in einem zweiseitigen Test analog über das entsprechende Konfidenzintervall für den interessierenden Parameter ϑ gefällt werden

- Die Nullhypothese H_0, also die Annahme $\vartheta = \vartheta_0$, ist beizubehalten, wenn der Wert ϑ_0 im $(1 - \alpha)$-Konfidenzintervall für den Parameter ϑ liegt

Allgemein bildet das $(1 - \alpha)$-Konfidenzintervall für den Parameter ϑ gerade den Beibehaltungsbereich der Nullhypothese

Diese Äquivalenz folgt aus der Identität der Teststatistiken mit den zur Ableitung der Konfidenzintervalle benutzten Pivot-Variablen (s. Besprechung spezieller Hypothesentests)

Multiple Tests

In empirischen Studien werden zur Beantwortung der wissenschaftlichen Fragen oftmals mehrere Hypothesentests durchgeführt

Angenommen, es werden k Tests durchgeführt, für die jeweils die Irrtumswahrscheinlichkeit α festgesetzt wird

Für diese Situation betrachten wir dann die Irrtumswahrscheinlichkeit α^*, dass bei den k Tests mindestens einmal ein Fehler 1. Art begangen wird (*familywise error rate*)

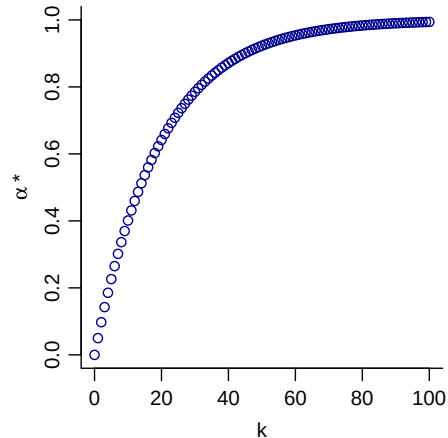
Sind die k Tests stochastisch unabhängig, dann gilt für die Wahrscheinlichkeit des Komplementärereignisses, dass in keinem Fall ein Fehler 1. Art begangen wird

$$(1-\alpha)^k$$

Familywise error rate

Damit ergibt sich für die Wahrscheinlichkeit α^* in mindestens einem Fall die Nullhypothese fälschlicherweise abzulehnen

$$\alpha^* = 1 - (1 - \alpha)^k$$



Bonferroni-Korrektur

Um eine derartige Inflation des Fehlers 1. Art bei multiplen Tests zu vermeiden, wurden verschiedene Korrekturen vorgeschlagen

Eine der am häufigsten verwendeten Korrekturen ist die *Bonferroni-Korrektur*

- Für k beliebige (nicht notwendig stochastisch unabhängige) Tests mit einer Irrtumswahrscheinlichkeit $\alpha_1, \ldots, \alpha_k$ gilt die Bonferroni-Ungleichung

$$\alpha^* \leq \alpha_1 + \ldots + \alpha_k$$

- Für identische Irrtumswahrscheinlichkeiten $\alpha = \alpha_i$, $i = 1, \ldots, k$ gilt dann

$$\alpha^* \leq k \cdot \alpha$$

Bonferroni-Korrektur

Will man sicherstellen, dass die Irrtumswahrscheinlichkeit α^* (familywise error rate) einen bestimmten Betrag nicht übersteigt (z. B. den Wert $\alpha^* \leq 0.05$), so sind die Irrtumswahrscheinlichkeiten der einzelnen Tests α entsprechend zu korrigieren

Aus der obigen Ungleichung ergibt sich dann die als Bonferroni-Korrektur bekannte Adjustierung

$$\alpha = \frac{\alpha^*}{k}$$

für die Irrtumswahrscheinlichkeit der einzelnen Tests

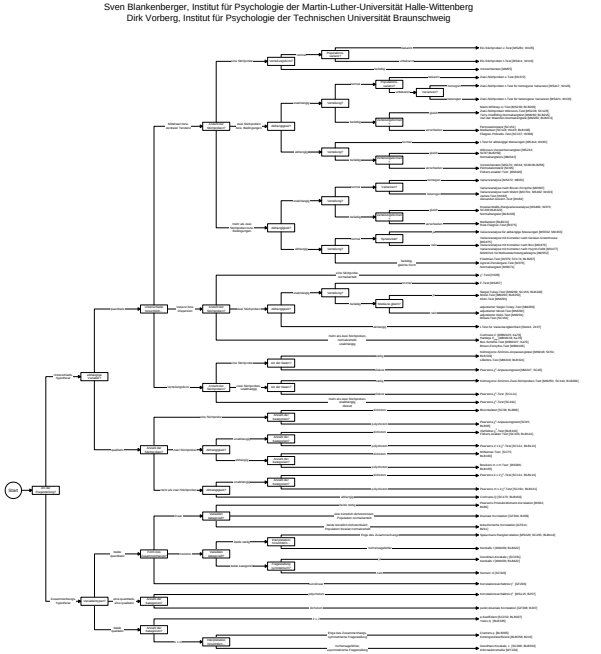
Beispiel: Bei 10 Tests, die mit einer Irrtumswahrscheinlichkeit von $\alpha^* \leq 0.05$ durchgeführt werden sollen, ist für die einzelnen Tests zu wählen

$$\alpha = \frac{\alpha^*}{10} = 0.005$$

Statistische Hypothesentests: Überblick

Statistische Tests lassen sich nach verschiedenen Gesichtspunkten klassifizieren, die sich auf die in den Test eingehenden Annahmen beziehen

- Skalenniveau der Ziel- und Einflussgrößen
- Annahmen zur Verteilung der betrachteten Zufallsvariablen
  - *Parametrische Tests* (*Parametertests*) gründen sich auf explizite oder approximative Verteilungsannahmen (v. a. Normalverteilung, auch über zentralen Grenzwertsatz)
  - *Nonparametrische Tests* verzichten auf Verteilungsannahmen
- Anzahl der betrachteten Merkmale bzw. Stichproben
  - Eine Stichprobe
  - Zwei (oder mehrere) Stichproben, die voneinander unabhängig oder abhängig sein können
- ...



Blankenberger & Vorberg (1998) präsentieren eine komplizierte Klassifikation statistischer Tests als Entscheidungsbaum

Lindeløv (2019) stellt dem die integrative Sicht entgegen, dass viele Tests nur Spezialfälle (verallgemeinerter) linearer Modelle sind, `https://lindeloev.github.io/tests-as-linear/`